USAID
FROM THE AMERICAN PEOPLE

EARLY GRADE READING ASSESSMENT BASELINE REPORT

AZAD JAMMU AND KASHMIR

SEPTEMBER 2014

This publication was produced for review by the United States Agency for International Development by. It was prepared by Management Systems International (MSI) with School-to-School International (STS) under the Monitoring and Evaluation Program (MEP).

EARLY GRADE READING ASSESSMENT BASELINE REPORT

AZAD JAMMU AND KASHMIR

Contracted under Order No. AID-391-C-00-13-00005

Monitoring and Evaluation Program (MEP)

DISCLAIMER

This study/report is made possible by the support of the American people through the United States Agency for International Development (USAID). The contents are the sole responsibility of Management Systems International and do not necessarily reflect the views of USAID or the United States Government.

ACKNOWLEDGEMENTS

We would like to thank the Education team of USAID/Pakistan for their forward planning to be able to collect baseline data before the roll out of the two important reading programs. Their support and responsiveness under a demanding timeline made this study possible. We would also like to thank the Azad Government of the State of Jammu and Kashmir Elementary and Secondary Schools Secretariat for their support of this activity. Finally, this effort would not have been possible without the dedication of our field teams of quality control officers and our local data collection partner, the Institute for Social and Applied Policy Studies.

CONTENTS

List of Tables and Figures

ACRONYMS

AJK	Azad Jammu and Kashmir
B.A.	Bachelor of Arts
B.Sc.	Bachelor of Science
C.T.	Certificate of Teaching (Grade 12 plus FA/FSC Certificate)
DEO	Data Entry Operators
DOE	Education Department
EGRA	Early Grade Reading Assessment
F.A.	Intermediate College (Grade 12) Certificate in Arts
FATA	Federally Administered Tribal Areas
F.Sc.	Intermediate College (Grade 12) Certificate in Sciences
GB	Gilgit-Baltistan
ICT	Islamabad Capital Territory
I-SAPS	Institute for Social and Applied Policy Studies
KP	Khyber Pakhtunkhwa
M.A.	Master of Arts
Matric	Secondary School (Grade 10) Certificate (Matriculation)
M.Ed.	Master of Education
M.Sc.	Master of Science
MSI	Management Systems International
MT	Master Trainers
NEAS	National Education Assessment System
NEMIS	National Education Management Information System
PRP	Pakistan Reading Project
P.T.C.	Primary Teaching (Grade 12) Certificate
QCO	Quality Control Officer
SPSS	Statistical Package for the Social Sciences
SRP	Sindh Reading Project
STS	School-to-School International
USAID	United States Agency for International Development

EXECUTIVE SUMMARY

Overview

In 2013, Management Systems International (MSI) and School-to-School International (STS) conducted a baseline reading assessment for primary school children prior to the launching of two USAID-funded projects: the Pakistan Reading Project (PRP) and the Sindh Reading Program (SRP). PRP is targeting improved reading for 910,000 children in Azad Jammu and Kashmir (AJK), Balochistan, the Federally Administered Tribal Areas (FATA), Gilgit-Baltistan (GB), the Islamabad Capital Territory (ICT), Khyber Pakhtunkhwa (KP), and Sindh, while SRP is targeting improved reading and mathematics for 750,000 children in Sindh. Targets will be achieved through support for 1) improved policies, laws, and guidelines for teachers and administrators, and 2) improved reading instruction for children in the primary grades.

To measure results from PRP and SRP, a rigorous external evaluation is being conducted. This report covers the baseline assessment in Azad Jammu and Kashmir. In May 2013, AJK, along with GB and ICT, was part of Round 1 of the baseline data collection; data from Pakistan's other five provinces/areas/territories (hereafter referred to as provinces) were collected in Rounds 2 and 3 in September and October 2013, respectively. The following activities were carried out for all of the provinces, including AJK: 1) design, 2) sampling, 3) instrumentation, 4) planning, 5) training, 6) implementation, 7) analysis, and 8) reporting.

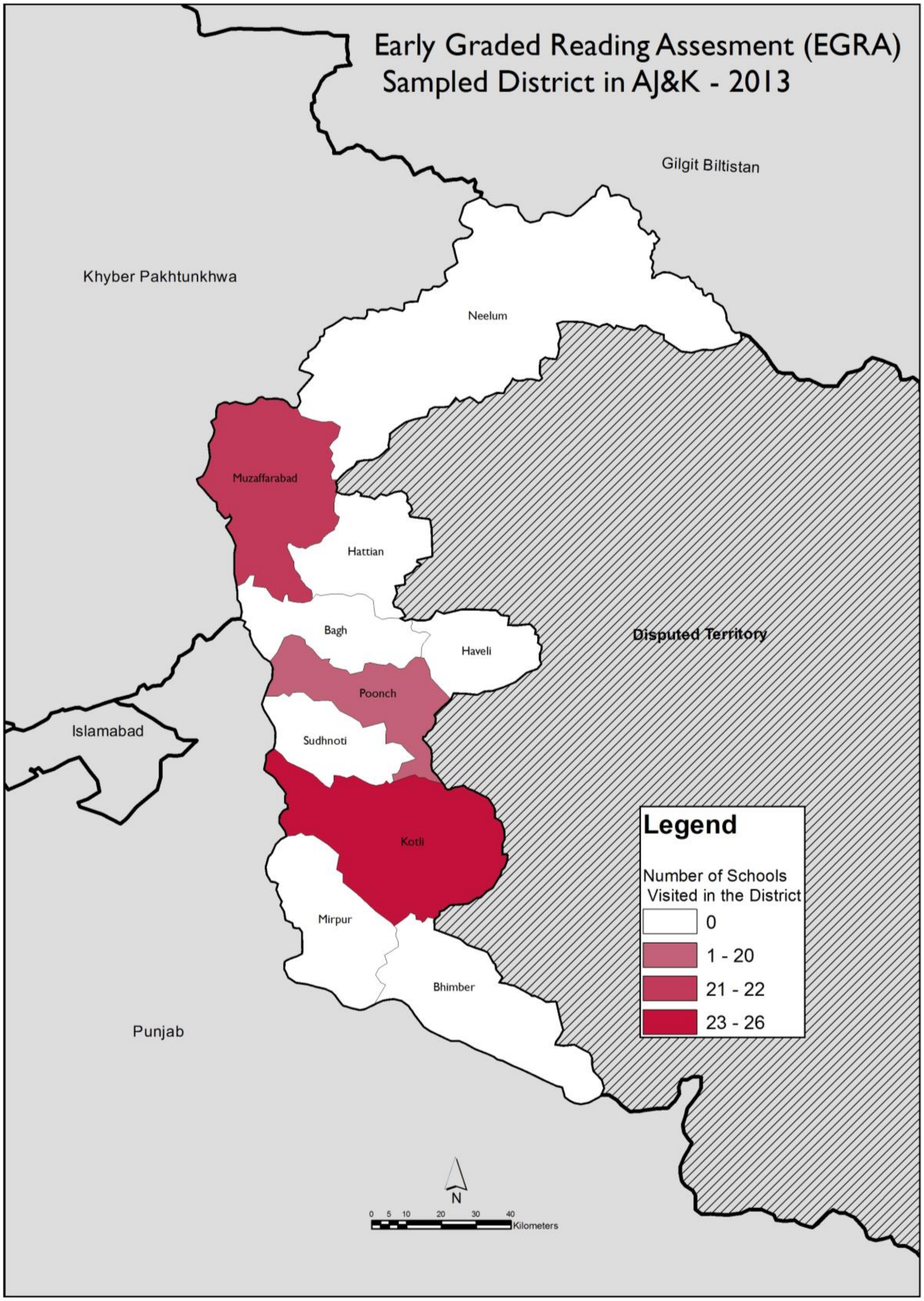
The external evaluation design, which was developed prior to the baseline assessment, was tailored to the implementation of the PRP and SRP in each province. In most of the provinces, a quasi-experimental design will be used, with two treatment groups: "full treatment" and "light treatment." The full treatment group will receive support in two areas: 1) policy, laws, and guidelines, and 2) improved instruction. The light treatment group will only receive the first kind of support.

In accordance with the USAID evaluation guidelines, students at two selected grade levels – grades 3 and 5 – were assessed at three time points: baseline, midline, and endline. An internationally accepted assessment tool, the Early Grade Reading Assessment (EGRA), was individually administered to a target sample of 33,600 children in 1,120 schools throughout the country. Over the course of the projects, the evaluators will compare the baseline results with those at the midline and endline to examine success in improving children's reading levels in Pakistan. The sampling was designed so that each province could be evaluated independently.

The long-term goal of this evaluation is to compare each province's baseline results to its midline and endline results, rather than other province's results. There are too many confounding variables – languages, curricula, administration dates, etc., that could render province-to-province comparisons meaningless. Furthermore, the evaluation is designed to investigate reading performance of the full and light treatment groups across time: baseline, midline, and endline. The differences between treatments will be fully investigated later, given the baseline data as the starting point for comparisons. In-depth comparisons between the full and light treatment groups are not useful at this time; such comparisons at baseline could add some bias by facilitating competition between the two groups that could compromise the validity of the evaluation.

For the baseline in AJK, all activities were completed by the end of September 2013, including a draft report. The EGRA baseline results were presented and discussed at a consultative meeting in Islamabad in September 2013. Representatives from the provincial Education Department (DOE), USAID, PRP, and the contractors (MSI and STS) attended the consultation. Revisions were then made to this report based on the discussions between the stakeholders.

Map of Sampled Districts



Key Points

Several key points from the EGRA baseline assessment in AJK are highlighted below:

Implementation

1. All districts in AJK were selected for “full treatment” during the initial consultative meetings between the DOE and USAID in January 2013. In AJK, there will not be a comparison of groups to determine the effects of the full treatment above and beyond those of the “light treatment”; rather, the results at the baseline will be compared against those in the midline and endline for the full treatment group only.

2. The Urdu versions of the EGRA tools were piloted in Muzaffarabad (and in districts in other provinces) prior to finalization. These were the tools used in AJK. The EGRA tools, which have been administered in various forms in over 40 countries, were successfully adapted for use in Pakistan. These included individually administered reading tests for students, along with questionnaires for students, teachers, and head teachers.

3. A total of 70 schools were selected for the baseline.

4. The baseline data were collected in a random sample of three districts in AJK: Kotli, Muzaffarabad, and Poonch. Within these districts, a random sample of male and female schools were selected, followed by a random sample of grades 3 and 5 students within those schools. The number of schools in the districts and the apportioned number of samples from each district by gender is shown in Table 2.

5. The results from this representative sample are presented below as a generalized view of the reading levels for children in AJK. Please note that district comparisons are not possible because the districts were not evenly sampled; the number of sampled schools varied by district, and the sample sizes are limited for each district.

6. The EGRA testing window was May 2013. All sampled schools were visited and all sampled students assessed during this period.

7. The assessment tools were successfully administered in the schools in the three sample districts as follows (with a percentage of the target reached in parentheses): 70 schools (100.0 percent) to 1,811 students (86.2 percent), 123 teachers (87.9 percent), and 69 head teachers (98.6 percent).

8. The validity and reliability of the tools was acceptable. Validity was assured through the adaptation process, which involved 17 educationists from throughout the country who participated in a workshop in Islamabad, and the standardized administration of the tools in the three districts. Reliability was assured through the high quality of the assessment tools and the standardized administration of the tools in Balochistan. Reliability estimates (of internal consistency) were calculated using the coefficient alpha.

9. The data entry and data cleaning process followed international standards. All student data were entered twice into two separate databases. These databases were then compared, with a resulting discrepancy rate of less than 1 percent. All data were reconciled across the two databases and with the assessment booklets. A clean data file was produced for analysis.

10. In the analysis phase, scores were calculated in three ways: 1) percentage correct scores for the reading tasks, 2) average percentage correct (grand means) for reading summary scores, and 3) adjusted raw scores for the timed reading tasks. (The calculation of these scores is fully explained in the analysis section of this report.) These scores provide a comprehensive picture of student

performance. Contextual analysis of student, teacher, head teacher, and school characteristics was carried out using the summary scores.

Results

1. EGRA was administered to 875 grade 3 students and 936 grade 5 students. The reliability estimates were acceptable for both grades (alpha = 0.79 for grade 3 and 0.78 for grade 5), indicating that the items worked well in measuring reading constructs at each grade level.

2. The task and item statistics showed that the EGRA discriminates well between low- and high-achieving students in both grades. The task p-values for grade 3 provided a spread on the lower to lower-middle section of the difficulty range, while p-values for grade 5 were higher and covered the upper-lower half to the high-middle parts of the spectrum. All task scores at grades 3 and 5 had item-total correlations equal to or greater than 0.30, indicating good discrimination quality for these tasks. (Complete item statistics are listed in Annex 1.)

3. The children in AJK did relatively well on familiar word reading and passage reading (fluency). There was also substantial progression from grade 3 to grade 5 on some of the tasks and for the summary scores. On the other hand, they had difficulty with phonics-related tasks such as letter sound knowledge and non-word reading. Passage and listening comprehension were also areas of weakness.

4. Passage reading (fluency) was nearly 50 points higher in grade 5 than in grade 3. This difference shows that the reading levels in grade 3 are low, but that children can make substantial progress in the early grades if expectations are high enough and if they are provided with the opportunity to learn.

5. Students were timed on five tasks as they read words or passages. These tasks were categorized into phonics (letter name recognition, letter sound knowledge, and non-word reading) and reading-rate fluency (familiar word and passage reading). Students at both grades had lower phonics scores than reading-rate fluency scores. Moreover, gains from grade 3 to grade 5 were lower for phonics than for reading-rate fluency tasks. Although the passage was designed for grade 3, this difference shows that the reading-rate fluency levels in grade 3 are low, but that students can make substantial progress in the early grades if expectations are high enough and if they are provided with the opportunity to learn. Specifically, mastery of phonics and phonemic awareness should help the students become better overall readers. It is clear that these types of knowledge and skills are not receiving an appropriate emphasis in schools in AJK.

6. Female students had higher scores, in general, than did their male counterparts. Areas such as letter name recognition, familiar word reading, passage reading, and passage comprehension were areas of particular strength for the females over the males in AJK. Boys tended to perform better than girls only in orientation to print. The differences between females and males were generally greater at grade 5 than in grade 3.

7. Questionnaire findings (for students, teachers, and head teachers) were mostly inconclusive, due to small sample sizes and the lack of variation in the scores in relation to the student, teacher, and head teacher characteristics. For the students, one of the positive findings was that attending a grade at an appropriate age seemed to have a positive effect on reading outcomes. In terms of the home environment, the presence of reading materials and the availability of a reading companion had some effects on outcomes, though they were limited.

Evaluation Recommendations

Given the success of the baseline assessment in AJK (and in the other provinces), the methods used in 2013 should be repeated as much as possible for the midline and endline assessments in future years. This should be conducted as follows:

1. The EGRA instruments proved to be of high quality, and equivalent versions of those tools should be developed – through trans-adaptation, piloting, and revision – for the midline and endline assessments so that progress can be accurately measured over time.

2. The EGRA items and tasks had good discrimination (quality) values and covered the low-to-middle part of the difficulty range. At baseline, the reading scores were relatively low for both grades and show room for growth. In addition, histograms and box pots provided evidence that the tool is expected to measure higher levels of reading-rate fluency that are anticipated following project-led interventions. Therefore, the baseline data indicates that the EGRA is appropriate for measuring increases in reading ability at midline and endline.

3. The sampling was reasonable in terms of finding a balance between the resources available, the required sample size, and the geographic coverage. It should be maintained in the midline and endline, i.e., keep the same districts and schools along with the sampling methods at the school level.

4. Since students can make great gains in reading during the primary grades, it is essential that testing occur at a consistent point in the academic year. Midline and endline testing in AJK should occur in May, thus matching the baseline timeframe and standardizing the instructional time across the study.

5. The systems for field data collection should be replicated, with the same systems for recruitment and training for the master trainers (MTs), field supervisors, quality control officers (QCOs), and enumerators as used in the baseline.

6. The data entry system should continue to be used, with the same systems for recruitment and training of data entry supervisors and operators, along with implementation through networked computers, double data entry, and reconciliation of errors.

7. The analysis should follow the same procedures, with calculations of task scores, summary scores, and timed task scores. The baseline, midline, and endline scores should be comparable so that improvements in children's reading can be accurately examined.

8. Reading proficiency levels should be created to provide educators and other stakeholders with meaningful results. Most parents and educators better understand reading achievement in useful terms or levels, such as emerging, proficient, or advanced, rather than interpreting a percent-correct test score that may differ by test or reading passage difficulty. Education officials are encouraged to select specific EGRA scores to serve as levels of reading proficiency for both grades. Percent correct for each task, summary score, as well as fluency rates are recommended for this purpose. The baseline EGRA data can be used for establishing these reading proficiency levels.

9. Finally, it may be advisable to add items to the student, teacher, and head teacher questionnaires for collecting data on PRP- and SRP-supported interventions so that student scores can be correlated with these indicators.

CHAPTER 1: INTRODUCTION

The Pakistan Reading Project (PRP) and the Sindh Reading Program (SRP) are two five-year initiatives funded by USAID. The projects/programs will cover over 40,000 government schools in Pakistan's eight provinces/areas/territories (hereafter referred to as provinces). PRP is targeting improved reading for 910,000 children in AJK, Balochistan, FATA, GB, ICT, KP, and Sindh, while SRP is targeting improved reading and mathematics for 750,000 children in Sindh. Targets will be achieved through support for 1) improved policies, laws, and guidelines for teachers and educational administrators, and 2) improved reading instruction for children in primary grades. Some districts in Pakistan will receive both kinds of support, i.e., "full treatment," while others will receive only the policy support, i.e., "light treatment." All schools within districts will receive the same type of treatment.

To measure results from PRP and SRP, a rigorous external evaluation is being conducted. The evaluation baseline is taking place in 2013, prior to the launch of the reading interventions. In accordance with USAID program evaluation guidelines, children are often tested at the ends of grades 2 and 4. In AJK, samples of children towards the beginning of grades 3 and 5 were assessed due to the timing of the baseline. Children at the same grade levels and at the same time of the year will be assessed at the midline and endline time points to evaluate the success of the interventions, taking into account the two treatment groups.

This report covers Azad Jammu and Kashmir. Along with GB and ICT, AJK was part of the baseline data collection in May 2013; data from Pakistan's other five provinces were collected in September and October 2013. The following activities were planned for all of the provinces, including AJK:

1. Design – USAID required a cross-sectional design, i.e., assessing students at the same grade levels (grades 3 and 5) over the course of PRP and SRP. In most provinces, this was complemented by a quasi-experimental design with the two treatment groups. However, AJK only had full treatment districts.

2. Sampling – The sampling plan enabled the collection of student reading assessment data that were representative of the treatment groups, grade levels, gender, and urban/rural zones. Schools were selected from all of the districts. There were a total of nine districts that were eligible for the sampling, and the Kotli, Muzaffarabad, and Poonch districts were selected at random (i.e., all districts except for Neelum -- see an explanation in the sampling section below). Schools were then apportioned according to location (urban/rural) and gender (boys/girls). As there are very few urban schools in AJK, balance for the location variable was not possible because too few urban schools were in the representative sample. Therefore, it was not appropriate to fully investigate the EGRA differences between urban and rural schools in AJK.

3. Instrumentation – EGRA tools were developed, primarily with grade 3 content, in English, Sindhi, and Urdu. Questionnaires for teachers, head teachers, and students were developed based on previous instruments used in other countries and on content specific to Pakistan. Model EGRA instruments were trans-adapted, piloted, revised, and finalized for use in Pakistan.

4. Planning – A field administration plan was developed for the baseline administration that would ensure the reliability of the data collected. The plan specified the timeline, training, logistics, field activities, supervision, data entry, analysis, reporting, and quality control.

5. Training – Workshops were conducted to train all master trainers, supervisors, enumerators, and QCOs. Enumerators and supervisors were observed to ensure clear comprehension and skills adequate to implement the EGRA tools.

6. Implementation – The baseline survey was implemented according to the plan. It ensured that all of the field activities took place in a standardized manner, as verified by the QCOs. The fieldwork was followed by data entry and preparation of a clean data file.

7. Analysis – Data were analyzed using spreadsheet (Excel) and statistical (SPSS) software. Experienced statisticians/psychometricians conducted the analysis, produced data tables and graphs, and ensured quality control.

8. Reporting – Provincial-level reports were produced. A reporting template was developed according to guidelines from the USAID contract. These reports will be disseminated to the provincial education authorities.

This report is organized into four chapters: 1) introduction, 2) methodology, 3) findings and results, and 4) conclusions and recommendations. Annexes with item statistics, box plots for the timed tasks, and a possible process for establishing a reading proficiency threshold follow the chapters.

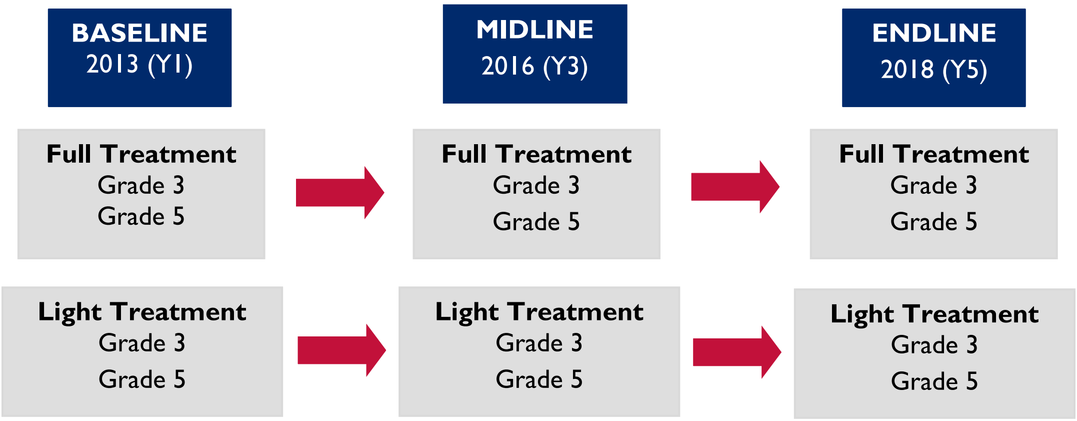
CHAPTER 2: DESIGN AND METHODOLOGY

This chapter presents the evaluation design and methodology, including the methods and systems used for collecting the EGRA baseline data. There are sections on the evaluation design, timeline, sampling, instrument development, data collection, data entry, and data analysis.

Evaluation Design

Following USAID policy, a cross-sectional evaluation design was developed prior to the baseline data collection. As shown in Figure 1, the design features two grade levels (3 and 5) and three time points (baseline, midline, and endline). Different groups of grade 3 and grade 5 students will be compared against each other across the three time points. In the figure, the years for the midline and endline are approximate and may be altered in accordance with implementation of the PRP and SRP interventions.

FIGURE 1: EVALUATION DESIGN



Districts for the "full" and "light" treatment groups – for those provinces with both groups – were pre-selected by the provincial DOEs and USAID during consultations in January and February 2013. Since district-level selection for the two groups was not random, equivalence at baseline of the two treatment groups cannot be assured, and a quasi-experimental design will be used. In this design, any differences in scores at baseline (and midline) will be statistically removed in the analysis, i.e., the two groups will be made statistically equivalent even though their average scores may be different. This will ensure fairness in the comparison of the full and light treatment groups.

In addition, while most districts have the two treatment groups, two of the provinces – AJK and ICT – will receive full treatment across all districts, and another province – FATA – will have full treatment in some districts, but no treatment (and no data collection) in the others. In AJK, all 10 of the districts will be covered by the PRP full treatment reading intervention. With this design, there will be no counterfactual (i.e., light treatment) for AJK.

In AJK, students were tested in Urdu, their main language of instruction. A few of the schools in AJK use English as the medium of instruction; however, the number of these schools was too small for the formation of a critical mass needed for evaluation purposes. Equal numbers of male and female schools were sampled for the EGRA testing; some mixed schools were included in the sample, but only boys or girls were selected from these schools, and thus they were considered as either male or female schools. Lists of the schools are safeguarded so they can be used again in the midline and endline data collections. The sampling design met the USAID requirements of adequate sample size and equal gender representation (see the sampling section below).

Timeline

The AJK baseline, like the other provinces, was conducted according to a timeline that started in January 2013 and continued through May 2014 with submissions of reports to USAID. The reports may then be distributed to the DOEs and other stakeholders as appropriate (see the timeline in Table 1).

The process began with the planning and design of activities, including the creation of preliminary sampling designs, selection of model EGRA tasks, recruitment of staff, and budgeting/contracting. This was followed in February by provincial consultations, including those for AJK. From February to April, the EGRA team, with participation from AJK and other provinces, then prepared, piloted, and revised the EGRA tools and conducted the district/school sampling. The data collection in AJK took place in May, and was followed by the data entry, analysis, and reporting from June to September, including the presentations to AJK DOE officials and USAID in September 2013.

TABLE 1: ROUND 1 TIMELINE (JANUARY 2013 TO MAY 2014)

Activity	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May
Plan and design EGRA activities	X	X															
Participate in provincial consultations	X	X															
Prepare EGRA tools		X	X														
Prepare test administration manuals			X														
Train master trainers and enumerators				X													
Select and verify sample schools			X	X													
Administer EGRA					X												
Enter data						X	X										
Analyze baseline data							X	X	X								
Produce draft reports								X	X								
Produce presentations									X								
Disseminate draft reports									X								
Make presentations									X								
Revise and finalize reports																	X
Submit reports to USAID																	X

Sampling

The sampling for Round 1 started in January 2013 with the selection of the treatment districts by the provincial DOEs and USAID. The EGRA team conducted the district and school sampling for Round 1, including AJK, in March and April. This included developing the sampling requirements, verifying the sample in the field, and finalizing the sample. The findings were provided in the sampling report for USAID.[1] As mentioned above, all districts in AJK were selected for full treatment. The sampling for AJK, as detailed in the sampling report, is briefly summarized in the following sub-sections of this report.

The sampling frame for AJK included nine of the 10 districts. During the consultation process, the DOE and USAID decided to intervene in all 10 districts, but to possibly exclude part of Neelum district due to security concerns. Since it was difficult to determine which schools in Neelum would be accessible to data collectors, the EGRA team eliminated that district from the sampling frame.

Sampling Requirements

Since the minimum requirement was 15 students per grade level in grades 3 and 5, only schools meeting that requirement were eligible for sampling. Within the selected districts and the eligible schools, equal numbers of male and female schools (35 each) were selected.

Sampling Process and Field Verification

Due to the need to balance representativeness, logistical demands, and resource availability (e.g., personnel, transportation, funding), one-third of the sampled population districts in AJK were chosen using a simple random sample. This resulted in a clustered sample. For the 35 male and 35 female schools (70 total – some of which were mixed schools), the samples were divided among the selected districts according to the proportions of schools within those districts (stratified random sampling). A second stratification was done at the "location" level, where schools were allocated by rural and urban according to their proportions in the National Education Management Information System (NEMIS); as seen in Table 2 below, there were relatively few urban schools in AJK (less than 2 percent). After sampling the 70 schools in AJK, an additional 10 male and 10 female schools were selected as replacements, if needed. Note that mixed schools may have been selected for some replacement schools due to not having enough options for replacement schools of strictly one gender. However, only students from the respective genders were included in those samples (i.e. if a mixed school was selected to replace a female school, only females were sampled).

TABLE 2: TARGET SAMPLE SCHOOLS BY DISTRICT, GENDER, AND LOCATION

District	Location	Schools	Pct.	Sample Schools		Replacement	
				Boys	Girls	Boys	Girls
Kotli	Rural	998	37	13	13	4	4
Kotli	Urban	35	1	0	0	0	0
Muzaffarabad	Rural	825	30	11	11	3	3
Muzaffarabad	Urban	64	2	0	0	0	0
Poonch	Rural	817	30	11	11	3	3
Poonch	Urban	10	0	0	0	0	0
Total		2,749	100	35	35	10	10

Once the schools were sampled, the QCOs, supplemented by EGRA senior managers, verified the samples in the field. This step was necessary due to two factors: 1) some inaccuracies in the NEMIS data and 2) changes in student numbers since the time period when the schools had submitted their data to

[1] MSI (2013). *Pakistan EGRA Sampling Report*. 18 June 2013 (Revised).

NEMIS. If the original schools had fewer than 15 students in either grade 3 or 5, a replacement school was selected and verified. At times, schools were retained if their numbers were near the minimum.

Intended and Actual Samples

When the random sample of districts was conducted, the districts selected were Kotli, Muzaffarabad, and Poonch. The number of schools in the districts and the apportioned number of samples from each district by gender and location is shown in Table 2 above.

After conducting the field verification, six schools – two male and four female – were replaced due to lower than expected numbers of students in the original samples. A total of 35 male and 35 female schools (including some mixed schools in each group), apportioned in the three districts and allocated according to the male/female and urban/rural requirements, were selected.

The actual number of students, teachers, and head teachers in the survey is presented in the results section.

Instrument Development

A brief summary of the instrument development process is presented below. The full results from the trans-adaptation, which involved educationists from AJK, were presented in a report to USAID.[2] This report is available to provincial education officials.

Trans-adaptation

In February, the EGRA team used tasks from the EGRA core instrument along with additional tasks used in instruments in other countries to develop a model test. Led by two international and two national assessment specialists, the EGRA team then organized a trans-adaptation workshop in Islamabad. A total of 17 English, Sindhi, and Urdu language specialists from the DOEs and teacher training institutes throughout Pakistan – including two subject specialists from AJK – participated in the workshop.

The trans-adaptation process involved the following with the local experts:

1. Discuss and choose reading tasks that would be of value to the baseline assessment in Pakistan;
2. Adapt each reading task using appropriate content in English, Urdu, and Sindhi; and
3. Ensure that the content would be suitable for grades 3 and 5 students.

The workshop resulted in a pilot EGRA test and pilot student, teacher, and head teacher questionnaires. The head teacher questionnaires included items about school characteristics.

Piloting

In March 2013, the EGRA English and Urdu tools were piloted in selected schools in AJK, ICT, and KP provinces, while the Sindhi tools were piloted in June in Sindh. Four tools were included in the pilot: 1) a student response booklet (including the student questionnaire), 2) a student stimuli booklet, 3) a teacher questionnaire, and 4) a head teacher questionnaire. The EGRA team conducted the pilot sampling, trained the enumerators, arranged the logistics, and supervised the piloting. The team then entered the pilot data into a database, analyzed the data, and developed preliminary recommendations for final tools in preparation for the revision workshop. They also prepared a piloting report for USAID.[3] As with the piloting report, the tools are available to provincial officials, though they must be kept secure since similar tasks will be used in the midline and endline.

[2] MSI (2013). *Pakistan EGRA Tools Trans-Adaptation Workshop Report*. June (Revised).

[3] MSI (2013). *Pakistan EGRA Instrument Development and Pilot Data Analysis*. August (Updated).

Revision and Finalization

The EGRA team held a revision workshop in March with a limited number of experts from the trans-adaptation workshop. Changes were made to the instruments based on the pilot data and field observations. These changes were summarized in the piloting report. The team then finalized the four instruments for each language and submitted them to USAID in April. USAID made suggestions, particularly around the inclusion of reading- and library-related items into the questionnaires that would provide baseline information for the PRP and SRP. The instruments were approved and then used in the training workshops in advance of the Round 1 data collection in May. The final instruments were comprised of the following:

- Students: 16 informational items; 8 tasks (one with 2 sub-tasks); and 34 questionnaire items
- Teachers: 15 informational items and 52 questionnaire items
- Head teachers: 17 informational items and 37 questionnaire items

These instruments are available for use by education officials.

Data Collection

Subcontractor Selection

The EGRA team, with the participation of USAID, issued a request for proposals and followed a set of criteria to select local subcontractors for the field data collection and for data entry. In April, the Institute for Social and Applied Policy Studies (I-SAPS) was chosen for both activities (data collection and data entry). MSI, STS, and I-SAPS collaborated on the data collection in AJK.

Data Collection

In April 2013, EGRA senior managers trained MTs and QCOs during a two-week session in Islamabad. The MTs then spent one week, also in Islamabad, training the I-SAPS AJK data collection team, which was comprised of a regional coordinator, two field supervisors, and 32 enumerators. The AJK team was trained alongside the teams from the other Round 1 provinces, i.e., GB and ICT. An EGRA senior manager and three QCOs were assigned to AJK to oversee and provide support for the I-SAPS team. The QCOs, coordinator, supervisors, and enumerators organized the logistics for the data collection. Following the training and logistical preparations in Islamabad, the QCOs and field supervisors conducted a two-day refresher course for the enumerators in Muzaffarabad just prior to commencing data collection in the schools.

Over a 10-day period in May, the enumerators spent a day in each of the 70 schools to collect the baseline data in AJK. The enumerators received frequent visits and mobile phone calls from the EGRA senior manager, QCOs, coordinator, and field supervisors to check on the status of data collection and to troubleshoot any issues. After collecting the data from the schools, the enumerators submitted their booklets to the supervisors and QCOs for verification and feedback. The supervisors then brought the booklets back to Islamabad for data entry.

Data Entry

Data Entry

In May 2013, the EGRA team developed a customized data entry application so that 1) the exact data from the booklets and questionnaires could be entered into a database, and 2) the computers used for data entry could be networked with a server. In June, the team trained the I-SAPS data coordinator, two supervisors, and 30 data entry operators (DEOs) on the application. In June and July, the EGRA and I-SAPS teams entered the data for nearly 10,000 student booklets, along with the questionnaires for the teachers and head teachers (Round 1). This included approximately 2,000 booklets and questionnaires for AJK.

Data Cleaning

In July, the EGRA and I-SAPS teams conducted the data verification and reconciliation. Following USAID requirements, 100 percent of the data were entered twice (double data entry) and any discrepancies between the first and second databases were reconciled. A clean data file was then provided to the data analysis team.

Data Analysis

Methodology

In June, the EGRA statisticians and psychometrician developed a research plan that included the following steps: 1) reliability estimates, 2) task and item statistics, 3) mean and grand mean scores (percent correct scores), 4) data plots, 5) timed and untimed task scores, and 6) questionnaire results. They used both SPSS and Excel for the analysis. Some of the analyses were replicated to ensure that the calculations were accurate. Descriptive analyses and inferential statistical comparisons were conducted by grade level and gender, and for the three sets of questionnaire data.

Please note that the analyses were only performed at the provincial level. This is because the sampling was conducted at the provincial level, i.e., the sample is only accurate at the provincial level. The samples at the district or school level are too small for analysis purposes, and any results at those levels would be misleading.

Validity and Reliability

Validity evidence for the tests was derived from previous experiences with EGRA in other developing countries, as well as through the trans-adaptation process in Pakistan. The test developers targeted grade 3 for the level of the tasks. An assumption was that the grade 5 students should perform better than the grade 3 students on each of the tasks.

For reliability, a generally accepted method is to estimate the internal consistency reliability (Coefficient Alpha) of the test. The minimum reliability threshold is approximately 0.75 to 0.80 for tests of this nature. Reliability was estimated for each province and language. Table 3 shows the reliability estimates for grades 3 and 5. These reliabilities are adequate and lend credibility to the internal consistency of the tests, indicating that the items are generally measuring similar reading constructs for both grade levels.

TABLE 3: RELIABILITY ESTIMATES

Language	Grade Level	Tasks	N-count	Alpha
Urdu	Grade 3	9	875	0.79
	Grade 5	9	936	0.78

Note that there were actually eight tasks, but one of the tasks (Task 7) was administered and scored in two parts, so the equivalent of nine tasks were used for the analysis.

Score Calculation

The EGRA data was analyzed three ways. First, p-values and item-total correlations were generated for assessing the difficulty and discrimination of the items and tasks. Second, the percent correct for each task provided an indication of the AJK students' mastery of the tasks, and third, AJK students' fluency was assessed.

Item P-values and Item-Total Correlations

P-values and item-total correlations are classical test theory statistics that are used to evaluate the performance of individual items and the tasks they comprise. Item difficulty is measured by p-values, which range from 0.00 to 1.00. Higher p-values indicate easier items, because a higher percentage of students posted correct responses. The other classical statistic is the item-total correlation, and it ranges from -1.00 to +1.00. This statistic measures how close the item or task relates to the overall percent correct on the summary score. Values above 0.2 are an indication of a good item or task.

Percent Correct

The results of the EGRA testing were calculated using task and summary scores. Table 4 lists the tasks, stimuli, raw score ranges, and the method for calculating the task and summary scores on the test. For each of the tasks, the stimuli (items) (i.e., questions, letters, sounds, words, and non-words) were worth one score point. The score points were added, and since the range of raw scores varies across the tasks, the percent of correct scores was used to report all results. No weighting was used with the tasks to calculate the summary scores. Each task summary score was calculated using the total number correct and dividing it by the number of items. The overall Reading Summary Score was calculated by adding all of the task summary scores and dividing by nine (total number of tasks) to arrive at the average.

Timed Tasks Scores

The scores on the timed tasks were calculated by taking the number of correct responses times 60 seconds then dividing that number by the number of seconds used to read the stimulus. For instance, if a student read 75 letters correctly in 30 seconds, their letters-correct-per-minute score would be 150 (75 words x 60 seconds/30 seconds). Given another example, if a student read 50 words correctly in 30 seconds, his or her timed task score would be 100 words per minute (50 words x 60 seconds/30 seconds). Table 4 lists the number of stimuli per task. Recall the percent correct scores ranged from zero to 100. The method for calculating phonics and fluency scores yielded much higher maximum values, upwards of 200 at baseline (see task box plots in Annex 2, Figures A1 and A2).

TABLE 4: EGRA SCORE RANGES AND CALCULATIONS

Task (Subtest)	Stimuli	Score Range	Calculation
1. Orientation to print	5 questions	0-5	Percent correct of answers
2. Letter name recognition	100 letters (timed)	0-100	Percent correct of letters
3. Phonemic awareness	10 questions	0-10	Percent correct of words
4. Letter sound knowledge	100 sounds (timed)	0-100	Percent correct of sounds
5. Familiar word reading	50 words (timed)	0-50	Percent correct of words
6. Non-word reading	50 non-words (timed)	0-50	Percent correct of non-words
7a. Passage reading	60 words (timed)	0-60	Percent correct of words
7b. Passage comprehension	5 questions	0-5	Percent correct of answers
8. Listening comprehension	3 questions	0-3	Percent correct of answers
Reading Summary Score	-	-	Average of percent correct

An example of percent correct scores for each of the tasks and as a summary score is provided below. The raw score is divided by the maximum score (the highest score possible in the score range) to produce the percent correct score for each task. Then, the task scores are averaged to produce the summary score. Note that each of the task percent correct scores is weighted equally to provide the summary score.

TABLE 5: EXAMPLE OF EGRA PERCENT CORRECT AND SUMMARY SCORES

Task (Subtest)	Maximum Score	Raw Score	% Correct Score
1. Orientation to print	5	3	60.0%
2. Letter name recognition	100	68	68.0%
3. Phonemic awareness	10	5	50.0%
4. Letter sound knowledge	100	42	42.0%
5. Familiar word reading	50	34	68.0%
6. Non-word reading	50	25	50.0%
7a. Passage reading	60	50	83.3%
7b. Passage comprehension	5	2	40.0%
8. Listening comprehension	3	1	33.3%
Reading Summary Score	--	--	55.0%

An example of timed task scores (adjusted) is provided below for the five fluency tasks. The formula explained above is used (timed task score = raw score x 60 seconds/seconds used).

TABLE 6: EXAMPLE OF EGRA TIMED TASK SCORES

Task (Subtest)	Raw Score	Seconds Used	Timed Task Score
2. Letter name recognition	68	48	85.0
4. Letter sound knowledge	42	60	42.0
5. Familiar word reading	34	48	42.5
6. Non-word reading	25	40	37.5
7a. Passage reading	50	40	75.0

CHAPTER 3: FINDINGS AND RESULTS

This chapter presents the findings and results from the EGRA baseline in AJK. There are sections on the sample, task and item statistics, score calculation, task and summary scores, timed task scores, and questionnaire findings.

Student Sample

The intended sample was 70 schools. Within these schools, the target was to assess 15 students in each grade, totaling 2,100 students (i.e., 1,050 for each gender and grade). Table 7 shows the number of students in the sample by grade and gender. For grade 3, the actual sample was 83.3 percent of the intended sample (n = 875); for grade 5, the total was 89.1 percent (n = 936).

A small number of students in grade 3 (n = 4) and grade 5 (n = 3) did not complete the gender item on the questionnaire. When analyzing the students by gender, the sample was 1,804 students (85.6 percent of the intended). However, when the data were not analyzed by gender, the total actual sample was 1,811 (86.2 percent of the intended).

During the field verification in April, most of the schools reported having at least the minimum number of 15 students in each of grades 3 and 5; a few schools were kept in the sample even though, during the field verification, their actual numbers were below the target. The main reason, however, for the difference between the intended and actual samples was low student attendance on the survey date.

TABLE 7: ACTUAL STUDENT SAMPLE BY GRADE AND GENDER

Grade Level	Sample	Boys	Girls	Missing	Total
Grade 3	Students	424	447	4	875
	% of Target	80.8%	85.1%	--	83.3%
Grade 5	Students	477	456	3	936
	% of Target	90.9%	86.9%	--	89.1%
Total	Students	901	903	7	1,811
	% of Target	85.8%	86.0%	--	86.2%

Task and Item Statistics

Table 8 shows the statistics for the tasks for the AJK sample. For each task, two statistics are provided: p-values and item-total correlations. P-values indicate the average score of the students on each of the tasks, or the difficulty of the tasks for the students. Item-total correlations in the table, which are actually task-total correlations, indicate the degree to which the tasks can discriminate between low and high achieving students; this is an indicator of the quality of the items. P-values can range from 0.00 to 1.00, with higher values indicating easier items. Item-total correlations can range from -1.00 to +1.00, with values above +0.20 indicating that the item (or task) is of high quality.

In Table 8 below, the task p-values for grade 3 in AJK ranged from 0.10 to 0.53, thus providing a spread on the lower half of the difficulty spectrum. The p-values for grade 5 were higher, ranging from 0.20 to 0.80 or in the middle parts of the range. The level of difficulty for both grade levels was appropriate for this baseline measure because there will be enough room in the scale for capturing growth during the midline and endline assessments. For item-total correlations, a generally acceptable threshold is 0.20 and above. Two tasks at grade 3 were below the threshold, while all others at both grade levels were above the threshold. This indicates good quality for the tasks. Complete item statistics are provided in Annex 1 at the end of this report.

TABLE 8: TASK STATISTICS

Task (Subtest)	Grade 3		Grade 5	
	P-Value	Item-Total	P-Value	Item-Total
1. Orientation to print	0.53	0.13	0.60	0.23
2. Letter name recognition	0.39	0.56	0.51	0.45
3. Phonemic awareness	0.31	0.19	0.40	0.25
4. Letter sound knowledge	0.11	0.38	0.20	0.37
5. Familiar word reading	0.39	0.78	0.77	0.71
6. Non-word reading	0.13	0.71	0.37	0.63
7a. Passage reading	0.41	0.77	0.80	0.68
7b. Passage comprehension	0.10	0.65	0.44	0.63
8. Listening comprehension	0.20	0.29	0.37	0.32

Task and Summary Scores

The next part of the analysis involves plotting the scores. Histograms of the summary scores (Figures 2 and 3) show that the distributions are moving to the right from grade 3 to grade 5, which is strong evidence that the children are learning basic skills at the primary school level. As with the task and item statistics, it also shows that there is room for growth at each grade level, particularly at grade 3. The goal of the intervention is to see movement of the distributions to the right within the same grade level (i.e., grades 3 and 5) from the baseline to midline to endline.

FIGURE 2: GRADE 3 SUMMARY SCORES

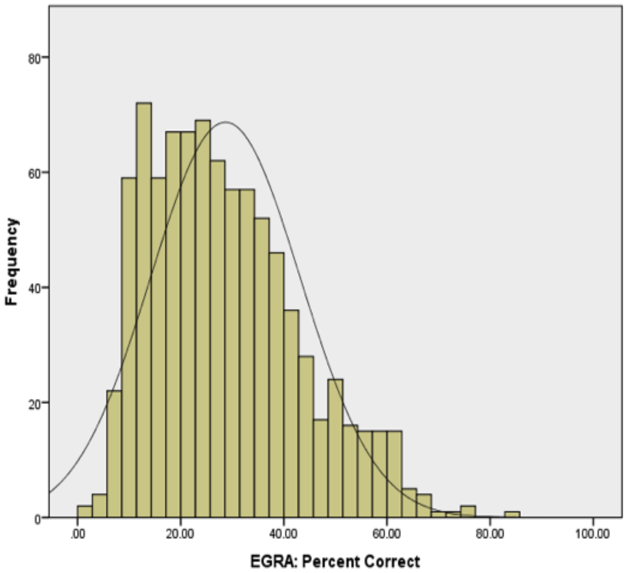


FIGURE 3: GRADE 5 SUMMARY SCORES

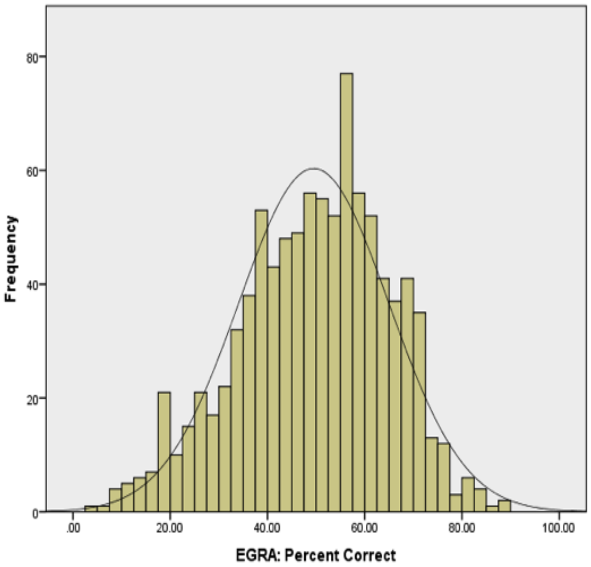


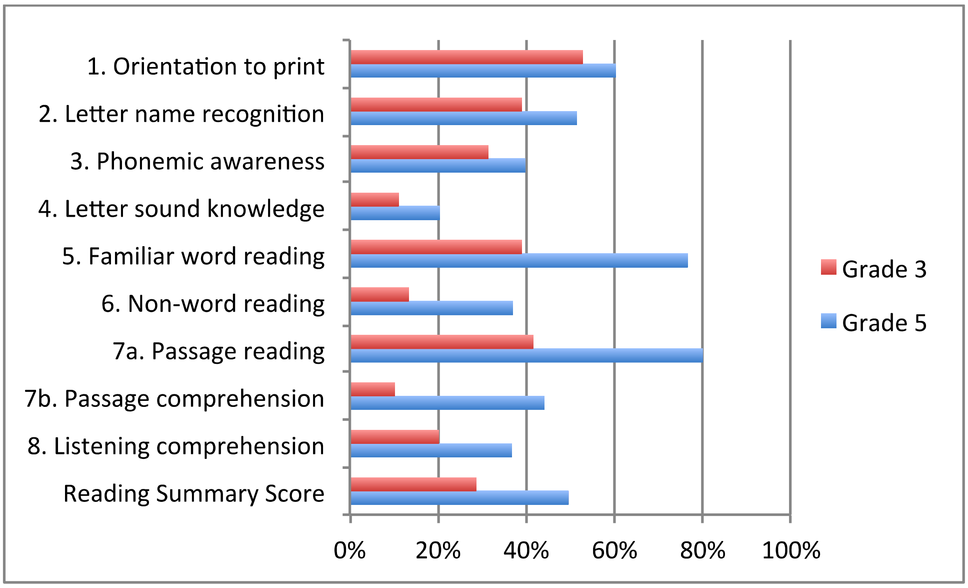
Table 9 and Figure 4 provide the average scores by task using percent correct scores. The score for each task was calculated using the total number correct and dividing by the number of items. For instance, a student who scored 3 out of 5 on Task 1 would receive a score of 60 percent. Averages were then calculated for all students on Task 1, which in AJK was 52.7 percent for grade 3 and 60.2 percent for grade 5. The same type of calculation was made for each student and each task. The table also includes the differences from grade 3 to grade 5, e.g., 60.2 percent minus 52.7 percent equals 7.5 percentage points.

Grade 3 students did relatively well in orientation to print, letter name recognition, familiar word reading, and passage reading, though their scores were under 50 percent in most of those areas. They had particularly low skills in areas such as letter sound knowledge, non-word reading, and passage comprehension. Grade 5 students showed strong increases in familiar word reading, passage reading, and passage comprehension. Phonemic awareness and letter sound knowledge are the two areas where the differences from grade 3 to grade 5 were the smallest and there is much room for improvement. In areas where there are large differences, interventions at grade 3 could have particularly large effects in accelerating children's learning.

TABLE 9: SCORES BY GRADE AND TASK

Task (Subtest)	Grade 3	Grade 5	Difference
1. Orientation to print	52.7%	60.2%	7.5% points
2. Letter name recognition	38.9%	51.4%	12.5% points
3. Phonemic awareness	31.3%	39.7%	8.4% points
4. Letter sound knowledge	11.0%	20.3%	9.3% points
5. Familiar word reading	38.9%	76.5%	37.6% points
6. Non-word reading	13.3%	36.9%	23.6% points
7a. Passage reading	41.5%	79.9%	38.4% points
7b. Passage comprehension	10.1%	44.0%	33.9% points
8. Listening comprehension	20.1%	36.7%	16.6% points
Reading Summary Score	28.6%	49.5%	20.9% points

FIGURE 4: SCORES BY GRADE AND TASK



When the scores were disaggregated by gender (Table 10 and Figures 5 and 6), most of the differences between boys and girls were small, though some were statistically significant ($p < 0.05$ level) in favor of girls. At grade 3, girls had higher scores in letter name recognition, passage reading, passage comprehension, and in the summary score. At grade 5, girls had higher scores in all areas except for orientation to print, letter sound knowledge, and listening comprehension. The differences in summary scores were 2.1 points and 6.1 points at grades 3 and 5, respectively.

TABLE 10: SCORES BY GRADE AND GENDER

Task (Subtest)	Grade 3		Grade 5	
	Boys	Girls	Boys	Girls
1. Orientation to print	54.0%	51.7%	61.6%	58.7%
2. Letter name recognition	37.0%	40.7%*	48.2%	54.6%*
3. Phonemic awareness	30.9%	31.7%	37.9%	41.4%*
4. Letter sound knowledge	10.9%	11.1%	19.0%	21.8%
5. Familiar word reading	36.6%	41.0%	71.3%	81.8%*
6. Non-word reading	12.5%	14.0%	31.8%	41.2%*
7a. Passage reading	38.3%	44.1%*	74.7%	85.2%*
7b. Passage comprehension	6.9%	12.9%*	37.1%	51.1%*
8. Listening comprehension	20.6%	19.6%	36.6%	36.8%
Reading Summary Score	27.5%	29.6%*	46.5%	52.6%*

* Indicates that the performance of the group was significantly higher, $p < 0.05$ level.

FIGURE 5: GRADE 3 SCORES BY TASK AND GENDER

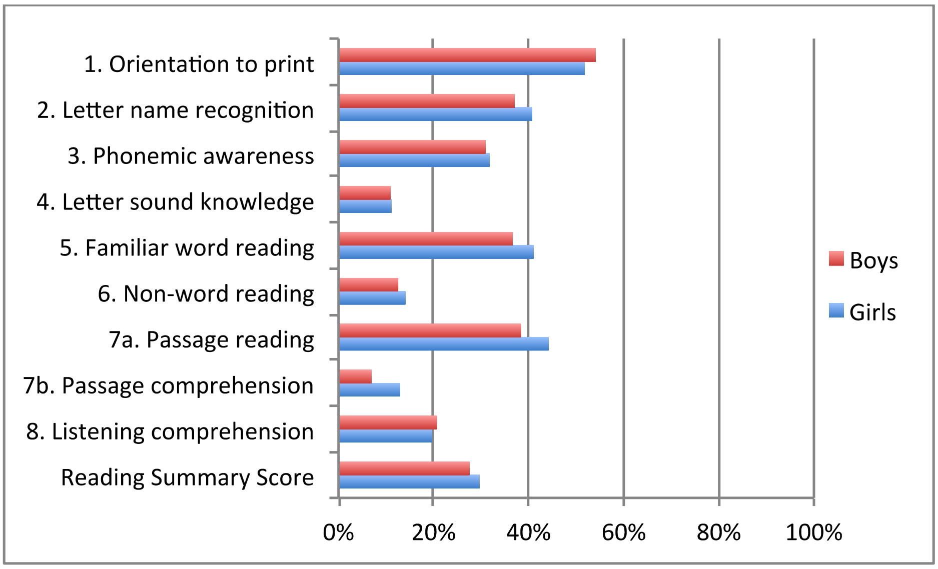
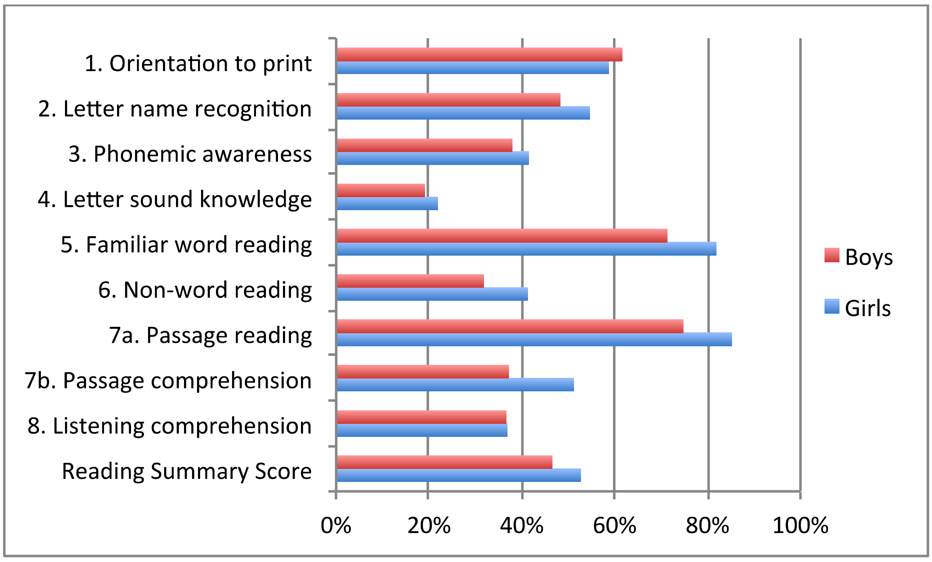


FIGURE 6: GRADE 5 SCORES BY TASK AND GENDER



Timed Tasks: Phonics and Reading-Rate Fluency Scores

Fluency is a measure of reading efficiency. On the Pakistan EGRA Baseline, there were two types of fluency measures: phonics and reading rate. The phonics-fluency subtest included letter name recognition, letter sound knowledge, and non-word reading, whereas, the reading-rate fluency subtest consisted of familiar word and passage reading.

Tables 9 to 11 below show scores in terms of raw scores (instead of the percent correct scores on the previous tables). Table 9 has the maximum raw scores attained by students on each task at each grade level. Tables 10 and 11 have mean scores for the students. In addition, adjustments were made to the raw scores for those students who finished the task before the end of one minute. For instance, if a student read 50 words correctly in 30 seconds, their words correct per minute score would be 100 (50 words x 60 seconds/30 seconds). Because these calculations are different from percent correct, the maximum scores are higher (see Figures A1 and A2 in Annex 2). Table 9 provides the baseline maximum scores at grade 3 and 5 for the five timed tasks.

TABLE 11: BASELINE MAXIMUM SCORES ON FLUENCY (TIMED) TASKS

Phonics Fluency Subtest	Grade 3	Grade 5
2. Letter name recognition	96	107
4. Letter sound knowledge	71	94
6. Non-word reading	60	80
Reading-Rate Fluency Subtest	**Grade 3**	**Grade 5**
5. Familiar word reading	118	140
7a. Passage reading	139	221

Table 12 shows the difference between grades, i.e., the progression from grade 3 to grade 5. The general term “points” was used to designate letters, sounds, words, or non-words. Table 13 shows the same scores by gender. For the timed tasks, passage reading had the most progression over the two grade levels. The lowest scores were in the areas of letter sound knowledge and non-word reading. These areas, along with letter name recognition, also showed the least progression from grade 3 to grade 5. By gender, there were significant differences in favor of girls at grade three on the letter name recognition, familiar word reading, and passage reading. These same differences were found at grade 5, with the addition of a difference in non-word reading. These timed task scores showed the same tendencies as the percent correct scores.

TABLE 12: TIMED TASK SCORES BY GRADE

Phonics Fluency Subtest	Grade 3	Grade 5	Difference (G5 – G3)
2. Letter name recognition	39.0	51.5	12.5 points
4. Letter sound knowledge	10.9	20.3	9.4 points
6. Non-word reading	7.2	20.1	12.9 points
Reading-Rate Fluency Subtest	**Grade 3**	**Grade 5**	**Difference (G5 – G3)**
5. Familiar word reading	21.7	54.8	33.1 points
7a. Passage reading	29.8	78.3	48.5 points

TABLE 13: TIMED TASK SCORES BY GRADE AND GENDER

Phonics Fluency Subtest	Grade 3		Grade 5	
	Boys	Girls	Boys	Girls
2. Letter name recognition	37.0	40.8*	48.3	54.7*
4. Letter sound knowledge	10.7	11.1	19.1	21.7
6. Non-word reading	7.1	7.1	17.0	23.3*
Reading-Rate Fluency Subtest	**Grade 3**		**Grade 5**	
	Boys	**Girls**	**Boys**	**Girls**
5. Familiar word reading	20.0	23.2*	48.1	61.7*
7a. Passage reading	26.7	32.4*	66.7	90.1*

*Indicates that the performance of the group was significantly higher, p< 0.01

Questionnaire Findings

Selected results are presented below, including for those characteristics or items that showed significant results. Note that there were some students, teachers, and head teachers who did not respond to certain questionnaire items; they were labeled as missing. The total averages for the summary scores were calculated based on those who responded. Statistical significance was determined based on *t*-tests for indicators with two categories and analysis of variance (with post hoc pairwise comparisons) for indicators with three or more categories.

Student Questionnaires

Table 14 shows the EGRA summary scores by student age. According to the National Education Policy (2009), the official age of the students at the beginning of the different grade levels of primary education is 6 to 10 years old. Since the baseline took place during the school year, the normal ages for this analysis were set at 8 to 9 years old for grade 3 and 10 to 11 years old for grade 5. The students were placed into three categories: younger than normal age for their grade, normal age, and older than normal age. There were significant differences in the scores. At grade 3, the highest scores were by the normal age students, while the younger students scored the highest at grade 5. The older students scored lower at each grade level. Post hoc comparisons between pairs of groups showed: 1) at grade 3, there was a significant difference between the normal and younger students; 2) at grade 5, there were significant differences between the younger and older students and between the normal and older students.

TABLE 14: SUMMARY SCORES BY STUDENT AGE

Age Group	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
Younger than normal age	80	24.2%	53	55.8%
Normal age	461	30.1%	414	51.7%
Older than normal age	333	27.7%	465	46.9%
Missing	1	--	4	--
Total	875	28.6%*	936	49.5%*

* Indicates that the performance of a group was significantly higher, p < 0.05 level.

Table 15 shows the EGRA summary scores by whether the student reads the Quran at home. While there are differences in the scores for each grade level, none of the differences were statistically significant, probably due to the small n-counts in some categories.

TABLE 15: SUMMARY SCORES BY READING THE QURAN AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	43	26.9%	31	44.5%
Yes	830	28.8%	902	49.7%
Missing	2	--	4	--
Total	875	28.6%	936	49.5%

* Indicates that the performance of the group was significantly higher, p < 0.05 level.

Table 16 shows the differences in scores based on whether there is a library at the school. While the results were statistically significant in favor of the students who said that there is a library at grade 3, the results were not significant at grade 5. Missing data on this item were higher than for most items.

TABLE 16: SUMMARY SCORES BY THE PRESENCE OF A LIBRARY AT THE SCHOOL

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	558	27.8%	553	49.9%
Yes	240	30.9%*	317	50.1%
Missing	77	--	66	--
Total	875	28.7%	936	50.0%

* Indicates that the performance of the group was significantly higher, p < 0.05 level.

In Tables 17 to 19, the data showed that the existence of newspapers and magazines (but not books) made a difference in reading scores in most cases. There may be evidence that increasing the presence of reading materials in the home could contribute to raising children's reading levels.

TABLE 17: SUMMARY SCORES BY THE PRESENCE OF NEWSPAPERS AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	652	27.9%	614	48.5%
Yes	223	30.8%*	322	51.4%*
Missing	0	--	0	--
Total	875	28.6%	936	49.5%

* Indicates that the performance of the group was significantly higher, p < 0.05 level.

TABLE 18: SUMMARY SCORES BY THE PRESENCE OF MAGAZINES AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	826	28.5%	771	48.4%
Yes	49	31.4%	165	54.8%*
Missing	0	--	0	--
Total	875	28.6%	936	49.5%

* Indicates that the performance of the group was significantly higher, p < 0.05 level.

TABLE 19: SUMMARY SCORES BY THE PRESENCE OF BOOKS AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	340	27.6%	388	50.3%
Yes	535	29.3%	548	48.9%
Missing	0	--	0	--
Total	875	28.6%	936	49.5%

* Indicates that the performance of the group was significantly higher, p < 0.05 level.

The final set of student questions (in Tables 20 to 22) pertained to children's reading habits at home. In general, these habits did not seem to make a difference in their scores, with the exception of grade 3 children practicing their reading to someone else.

TABLE 20: SUMMARY SCORES BY CHILDREN HAVING SOMEONE READ TO THEM AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	332	27.8%	274	49.1%
Yes	532	29.3%	657	49.7%
Missing	11	--	5	--
Total	875	28.6%	936	49.5%

* Indicates that the performance of the group was significantly higher, p < 0.05 level.

TABLE 21: SUMMARY SCORES BY CHILDREN READING TO SOMEONE ELSE AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	302	26.9%	261	49.5%
Yes	567	29.7%*	671	49.6%
Missing	6	--	4	--
Total	875	28.6%	936	49.5%

* Indicates that the performance of the group was significantly higher, $p < 0.05$ level.

TABLE 22: SUMMARY SCORES BY CHILDREN READING SILENTLY AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	219	29.1%	234	49.6%
Yes	644	28.6%	698	49.5%
Missing	12	--	4	--
Total	863	28.6%	932	49.5%

* Indicates that the performance of the group was significantly higher, $p < 0.05$ level.

Teacher Questionnaires

With the smaller sample size, the analysis of the teacher questionnaires was limited to providing descriptive statistics on teacher characteristics and summary scores, i.e., with no group comparisons. Tables 23 to 27 provide information on teacher academic qualifications, professional qualifications, age, years of experience, and in-service training.

TABLE 23: SUMMARY SCORES BY TEACHER ACADEMIC QUALIFICATION

Academic Qualification	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
M.A./M.Sc.	11	28.6%	12	47.1%
B.A./B.Sc.	17	27.5%	29	51.4%
F.A./F.Sc.	8	28.3%	6	47.6%
Matric	22	29.1%	15	48.6%
Missing	2	--	1	--
Total	60	28.4%	63	49.5%

TABLE 24: SUMMARY SCORES BY TEACHER PROFESSIONAL QUALIFICATION

Professional Qualification	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
M.Ed./M.A.	3	25.1%	3	50.4%
B.Ed.	19	28.7%	26	50.0%
C.T.	9	27.4%	10	50.9%
P.T.C.	21	28.6%	17	46.8%
Missing	8	--	7	--
Total	60	28.2%	63	49.2%

TABLE 25: SUMMARY SCORES BY TEACHER AGE

Age Group in Years	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
40 and less	24	27.2%	20	49.4%
Between 41 and 50	26	28.5%	29	49.1%
51 and more	9	29.3%	10	52.8%
Missing	1	--	4	--
Total	60	28.1%	63	49.8%

TABLE 26: SUMMARY SCORES BY TEACHER EXPERIENCE

Years of Experience	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
10 or less	23	27.3%	20	48.5%
Between 11 and 20	8	29.7%	7	50.0%
Between 21 and 30	22	29.1%	29	49.8%
31 or more	7	27.6%	7	51.8%
Missing	0	--	0	--
Total	60	28.3%	63	49.6%

TABLE 27: SUMMARY SCORES BY TEACHER IN-SERVICE TRAINING

Frequency of Training	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
None	47	28.4%	46	48.7%
One time	10	28.3%	11	49.9%
Two times	3	25.8%	2	59.6%
Three times	0	--	3	57.3%
Missing	0	--	1	--
Total	60	28.3%	63	49.7%

While there were some small differences in student summary scores by teacher characteristic on some of the variables, no clear patterns emerged. Any observed differences should be treated with caution due to the small sample size.

Head Teacher Questionnaires

Similar to the teacher questionnaires, the sample size for the head teacher questionnaires was small, so interpretations of the data should be treated with caution. The characteristics presented are head teacher academic qualification, professional qualification, experience, in-service training, support to teachers in reading, and training in supporting reading (Tables 28 to 33). As with the teacher data, no clear conclusions can be found with the head teacher characteristics and student scores.

TABLE 28: SUMMARY SCORES BY HEAD TEACHER ACADEMIC QUALIFICATION

Academic Qualification	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
M.A./M.Sc.	43	30.7%	43	51.5%
B.A./B.Sc.	22	24.4%	22	46.9%
F.A./F.Sc.	1	24.4%	1	44.9%
Matric	2	35.6%	2	57.4%
Missing	1	--	1	--
Total	69	28.7%	69	50.0%

TABLE 29: SUMMARY SCORES BY HEAD TEACHER PROFESSIONAL QUALIFICATION

Professional Qualification	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
M.Ed./M.A.	18	29.2%	18	50.3%
B.Ed.	46	28.5%	46	50.1%
C.T.	3	23.6%	3	41.6%
P.T.C.	2	35.6%	2	57.4%
Missing	0	--	0	--
Total	69	28.7%	69	50.0%

TABLE 30: SUMMARY SCORES BY HEAD TEACHER EXPERIENCE

Years of Experience	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
2 or less	11	27.8%	11	50.3%
3 to 5	19	27.4%	19	49.6%
6 to 10	13	31.6%	13	50.0%
11 or more	25	28.5%	25	50.3%
Missing	1	--	1	--
Total	69	28.7%	69	50.0%

TABLE 31: SUMMARY SCORES BY HEAD TEACHER IN-SERVICE TRAINING

Frequency of Training	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
None	59	28.4%	59	49.4%
1 time	6	33.1%	6	56.6%
2 times	2	34.6%	2	56.0%
More than 2 times	1	16.3%	1	33.7%
Missing	1	--	1	--
Total	69	28.8%	69	50.0%

TABLE 32: SUMMARY SCORES BY HEAD TEACHER SUPPORT TO TEACHERS IN READING

Support to Teachers	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	41	27.9%	41	48.3%
Yes	26	30.3%	26	52.4%
Missing	2	--	2	--
Total	69	28.8%	69	49.9%

TABLE 33: SUMMARY SCORES BY HEAD TEACHER TRAINING IN TEACHING READING

Support to Teachers	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	3	31.7%	3	55.5%
Yes	66	28.5%	66	49.8%
Missing	0	--	0	--
Total	69	28.7%	69	50.0%

School Characteristics

The final section provides information on school characteristics (from the head teacher questionnaires) by student summary scores. As with the teacher and head teacher characteristics, the sample size of the school characteristics was too small for any statistical comparisons (Tables 34 to 37). Any patterns were inconclusive.

TABLE 34: SUMMARY SCORES BY SCHOOL GENDER

School Gender	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
Male school	16	26.7%	16	46.5%
Female school	24	29.3%	24	53.7%
Mixed school	28	28.8%	28	48.8%
Missing	2	--	2	--
Total	70	28.5%		50.0%

TABLE 35: SUMMARY SCORES BY PTA/SMC/PTSMC/PTC

Parent Teacher Committee	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	21	26.5%	21	50.7%
Yes	45	30.5%	45	50.5%
Missing	4	--	4	--
Total	70	29.2%	70	50.6%

TABLE 36: SUMMARY SCORES BY PRESENCE OF A SCHOOL LIBRARY

School Library	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	53	28.6%	53	49.6%
Yes	15	27.9%	15	51.4%
Missing	2	--	2	--
Total	70	28.4%	70	50.0%

TABLE 37: SUMMARY SCORES BY INFRASTRUCTURE (DRINKING WATER, ELECTRICITY, TOILETS)

Number of Infrastructures (Water, Electricity, Toilets)	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
None	11	30.0%	11	51.0%
1	20	26.2%	20	47.2%
2	20	29.3%	20	49.7%
3	19	29.5%	19	52.4%
Missing	0	--	0	--
Total	70	28.6%	70	49.9%

CHAPTER 4: CONCLUSIONS AND RECOMMENDATIONS

This final chapter provides conclusions from the AJK EGRA baseline. It is organized according to the two main sections in the report: 1) design and methodology, and 2) findings and results. There are also recommendations based on the instrument development, data collection, data entry, and analysis.

Design and Methodology

1. The design followed USAID evaluation guidelines for a cross-sectional approach. However, due to selecting all of the districts (and all of the schools) in AJK for full treatment, there is no counterfactual against which to measure the effects of the full treatment above and beyond the light treatment. With the cross-sectional design, the evaluation will be limited to examining the progress of children in grades 3 and 5 over the course of the PRP project.
2. The sampling issues were addressed as well as could have been expected. The main issue was the lack of schools with the requisite number of children per grade level. However, the actual sample of schools was 100 percent and the actual sample of students reached over 87 percent of the intended sample.
3. The EGRA test was of good quality. The task statistics were acceptable, with an appropriate range of p-values and item-total correlations that were at an acceptable level of quality. The characteristics of the test were such that it should be a strong measure of progress over time due to project-led interventions. As with any test, there may be ways to improve on the task and item statistics for the midline and endline.
4. The field implementation was successful, though there were difficulties to overcome, including logistical challenges with the difficult terrain in AJK. In spite of these challenges, there was a high level of standardization reported by the quality control officers, which they attributed to the effective training process by the EGRA team. The team paid careful attention to detail in the logistics and test administration, which was reflected in the low error rates in the booklets and in the data entry.

Findings and Results

The AJK evaluation involves one kind of analysis: a comparison of each group to itself at the baseline, midline, and endline. There is no counterfactual in AJK since all of the districts will receive full treatment and all districts (except for Neelum) were eligible for sampling. Please see Figure 1 and the accompanying text for a fuller description of the evaluation design.

Several key findings emerged from the baseline assessment in AJK. These are as follows:

1. EGRA was administered to 875 grade 3 students and 936 grade 5 students. The reliability was acceptable for both grades (alpha = 0.79 and 0.78 for grades 3 and 5, respectively). These levels of reliability indicate that the items worked well in measuring reading constructs at both grade levels. The task and item statistics showed that the EGRA discriminates between low- and high-achieving students in both grades. The task p-values for grade 3 provided a spread on the lower to lower-middle section of the difficulty range, while p-values for grade 5 were higher, covering the upper-lower half to the high-middle parts of the spectrum. All task scores at grades 3 and 5 had item-total correlations equal to or greater than 0.30, indicating good discrimination quality for these tasks. (Complete item statistics are listed in Annex 1.)
2. Students had the most difficulty with phonics-related tasks such as letter sound knowledge and non-word reading. Passage and listening comprehension were also areas of weakness. On the other hand, students did well on familiar word reading and passage reading (fluency). There was also substantial progression from grade 3 to grade 5.

3. Passage reading (fluency) was nearly 50 points higher in grade 5 than in grade 3. Although the passage was designed for the third grade level, this difference shows that the reading levels in grade 3 are low, but that students can make substantial progress in the early grades if expectations are high enough and if given the opportunity. Specifically, mastery of phonics, such as letter sound knowledge and non-word reading, should help the students become better overall readers. It is clear that these types of knowledge and skills are not receiving an appropriate emphasis in schools in AJK.

4. Female students had higher scores, in general, than their male counterparts. Areas such as letter name recognition, familiar word reading, passage reading, and passage comprehension were areas of particular strength for the females over the males in AJK. The differences were greater at grade 5 than in grade 3.

5. Questionnaire findings were mostly inconclusive, due to small sample sizes and the lack of differences in responses within the student, teacher, and head teacher samples. For the students, attending a grade at an appropriate age seemed to have a positive effect on reading outcomes. In terms of the home environment, the presence of reading materials and a person to read to appeared to have some effect on outcomes, though it was limited.

Evaluation Recommendations

Given the success of the baseline assessment in AJK (and in the other provinces), the methods used in 2013 should be repeated as much as possible for the midline and endline assessments in future years. This should be conducted as follows:

1. The EGRA instrument development and trans-adaptation process was comprehensive and resulted in high quality EGRA tools. This should be repeated as soon as possible with the tasks that need to be changed for the midline and endline tools (to minimize test-retest effects and security breaches), so that reading progress can be accurately measured over time.

2. The EGRA items and tasks had good reliability values and covered the low-to-middle difficulty range. At baseline, the reading scores were relatively low for both grades, and show room for growth. In addition, histograms and box pots provided evidence that the tool is expected to measure higher levels of reading that are anticipated due to project-led interventions. Therefore, the baseline data indicates that EGRA is appropriate for measuring increases in reading ability at midline and endline.

3. The sampling was reasonable in terms of finding a balance between the resources available, the required sample size, and the geographic coverage. It should be maintained in the midline and endline, i.e., keep the same districts and schools, along with the methods at the school level.

4. The systems developed for field data collection should be repeated. The different layers of management, coordination, supervision, and quality control contributed to successful planning, implementation, and problem solving. The quality control officers were particularly important in maintaining standards and providing support for the local subcontractors.

5. The data entry process took time to develop but it eventually proved to be advantageous in terms of having the data entry operators connect to a central server. This facilitated the two rounds of data entry and the reconciliation process. This system should also be repeated in subsequent data entry activities.

6. Since AJK was part of Round 1, the methods for analysis also took some time to develop, but it was important to create templates and agree on a methodology due to the volume of analysis and reporting for the eight provinces. Again, the investment of time and effort in this process was needed for Rounds 2 and 3 of the baseline and for the midline and endline.

7. In terms of the technical properties of the assessment, there may be some merit in providing a reading passage for the grade 5 students that is at their level. Their performance on the fluency task was relatively high, showing that many of the grade 5 students can read grade 3 material. A grade 5 reading passage might be a more appropriate measure for these students. The results, if reported by performance levels, would provide data for determining whether the grade 5 students are reading fluently at grade level. (Note that comprehension for these students remains a problem, i.e., even for the questions developed for a student at grade 3.)

In general, the AJK baseline was successful in providing accurate data on which to base decisions for implementation of the PRP interventions, and also for tracking student reading progress over time. It provides a solid foundation for the midline and endline assessments.

ANNEXES

Annexes 1 to 4 provide additional information on the EGRA baseline. Specifically, the annexes have the following:

Annex 1 gives complete item statistics – p-values (the difficulty of the items) and item-total correlations (the quality of the items) by grade – for the items associated with the various tasks. These are more detailed than the task statistics presented in Chapter 3 of the report. Measurement specialists often request these kinds of item statistics for the purposes of quality control, analysis, and test equating.

Annex 2 provides box plots for the fluency tasks. The box plots are more task-specific than the overall score distributions (histograms) presented in the report. They show the median (middle score), the range (highest and lowest scores), and the distribution of scores (by quartiles) for each task. The task-specific distributions are useful to EGRA specialists who place emphasis on the fluency tasks.

Annex 3 gives two examples of categorizing passage reading fluency scores using performance levels. The categorizations – along with raw scores and scale scores -- are often used to interpret test scores. The first example combines reading speed with comprehension, while the second example only uses reading speed. Each example uses a set of cut-scores for placing the students into performance categories.

Annex 4 provides detailed information on the second example, with results for each category of fluency and each level of comprehension. These data can be used as evidence on the reliability of using a combined measure of fluency and comprehension for setting performance cut-scores. The validity of combining these scores is more of an issue for reading experts.

Annex 1: Complete Item Statistics by Grade

Table A1 presents item statistics for the untimed tasks, each of which have multiple items. For instance, task 1 (orientation to print) has item statistics for its five items (Q1 to Q5). Note that the timed tasks are lists of letters, sounds, and words, i.e., not items, so it is not necessary to calculate item statistics for them.

Previously, we presented task statistics (Chapter 3, Table 8) with explanations of how they are calculated. These item statistics are calculated in the same way. They show the difficulty and quality of the items. Recall that when constructing a test, we strive for tasks and items that have difficulty values (p-values) that are spread across the range from about 0.05 to 0.90 and quality values (item-total correlations) of at least 0.20. The difficulty values ranged from 0.04 to 0.71 for grade 3 and 0.08 to 0.83 for grade 5, indicating a strong range of item difficulties. A total of 15 and 21 items for grades 3 and 5 respectively out of the 23 items per grade had item-total correlations of at least 0.20, indicating high quality items.

TABLE A1: COMPLETE ITEM STATISTICS BY GRADE

Task (Subtest)	Item	Grade 3		Grade 5	
		P-Value	Item-Total	P-Value	Item-Total
1. Orientation to print (untimed)	Q1	0.70	0.39	0.71	0.39
	Q2	0.71	0.43	0.78	0.40
	Q3	0.41	0.20	0.46	0.15
	Q4	0.12	0.09	0.23	0.08
	Q5	0.71	0.16	0.83	0.21
2. Letter name recognition (timed)	--				
3. Phonemic awareness (untimed)	Q1	0.52	0.27	0.67	0.40
	Q2	0.23	0.26	0.38	0.45
	Q3	0.28	0.19	0.30	0.28
	Q4	0.22	0.19	0.35	0.41
	Q5	0.33	0.17	0.37	0.39
	Q6	0.43	0.22	0.53	0.34
	Q7	0.20	0.26	0.25	0.37
	Q8	0.25	0.32	0.33	0.36
	Q9	0.17	0.25	0.23	0.37
	Q10	0.50	0.26	0.55	0.33
4. Letter sound knowledge (timed)	--				
5. Familiar word reading (timed)	--				
6. Non-word reading (timed)	--				
7a. Passage reading (timed)	--				
7b. Passage comprehension (untimed)	Q1	0.06	0.40	0.41	0.39
	Q2	0.12	0.44	0.39	0.32
	Q3	0.04	0.29	0.18	0.30
	Q4	0.13	0.48	0.60	0.41
	Q5	0.07	0.56	0.32	0.25
8. Listening comprehension (untimed)	Q1	0.25	0.13	0.44	0.22
	Q2	0.05	0.13	0.08	0.21
	Q3	0.31	0.16	0.58	0.21

Annex 2: Box Plots for Phonics and Reading-rate Fluency Tasks

EGRA places a high emphasis on fluency (timed) tasks. In addition to the descriptive statistics in Table 9 (percent correct scores) and Table 14 (fluency task means), we show box plots for the different fluency tasks. Widely used since their development in the 1960s, box plots are a convenient way for graphically presenting numerical data.

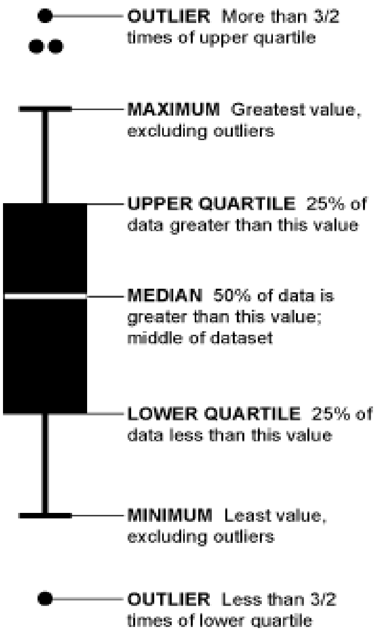
Box plots have two characteristics: 1) central tendency (i.e., the median, or the middle score in the data) and 2) variation (i.e., the range, with scores grouped by quartile). The boxes (which are actually rectangles) represent the two middle quartiles of the scores and the "whiskers" represent the upper and lower quartiles. The small circles on the ends of the whiskers represent outliers. The figure below provides a more detailed explanation for interpreting box plots.

FIGURE A1: UNDERSTANDING BOXPLOTS

Median
The median (middle quartile) marks the mid-point of the data and is shown by the line that divides the box into two parts. Half the scores are greater than or equal to this value and half are less.

Upper quartile
Seventy-five percent of the scores fall below the upper quartile.

Lower quartile
Twenty-five percent of scores fall below the lower quartile.


OUTLIER More than 3/2 times of upper quartile
MAXIMUM Greatest value, excluding outliers
UPPER QUARTILE 25% of data greater than this value
MEDIAN 50% of data is greater than this value; middle of dataset
LOWER QUARTILE 25% of data less than this value
MINIMUM Least value, excluding outliers
OUTLIER Less than 3/2 times of lower quartile

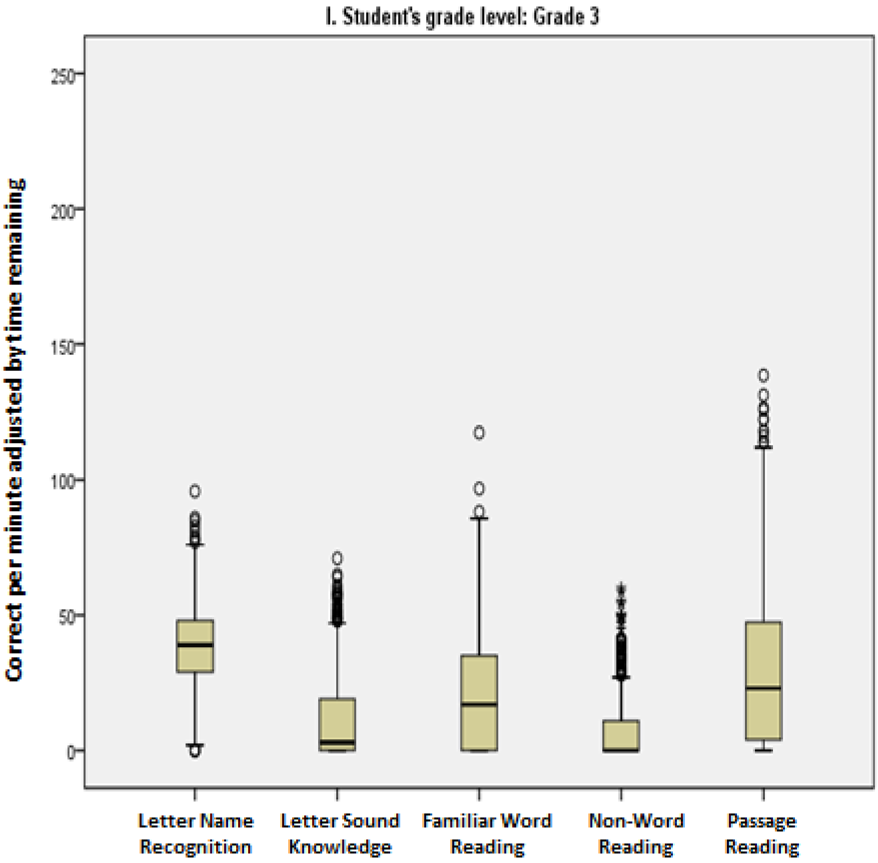
Box plots are presented below (Figures A2 and A3) for the results by grade level on the five fluency (timed) tasks: letter name recognition (task 2), letter sound knowledge (task 4), familiar word reading (task 5), non-word reading (task 6), and passage reading (task 7a).

Grade 3

For grade 3, the central tendency (i.e., the median speed, or the line in the middle) for each of the tasks ranged from about 0 (non-word reading) to about 40 (letter name recognition) items per minute. It shows that the students had much better knowledge of letter names than of their grapheme-phoneme correspondence.

The variation (i.e., the range of scores, without outliers) for each of the tasks varied from about 30 (non-word reading) to about 130 (passage reading). It shows that the scores were more spread out when reading connected words than sounding out pseudo-words.

FIGURE A2: PHONICS AND READING-RATE FLUENCY BOX PLOTS FOR GRADE 3



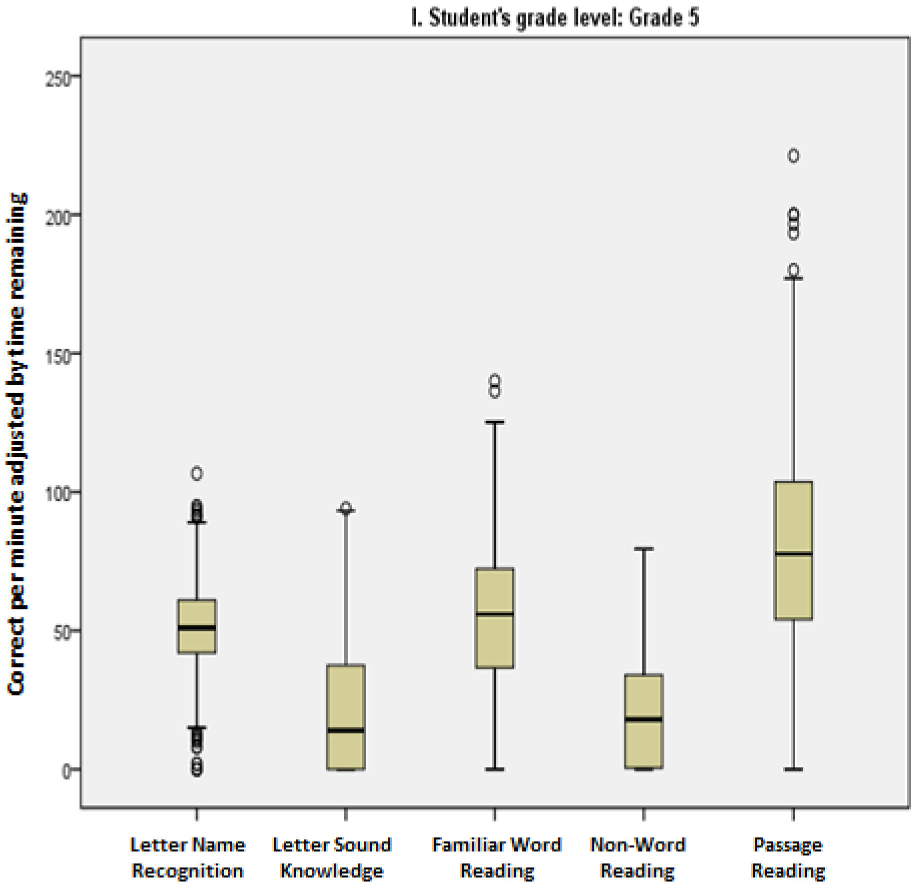
Grade 5

For grade 5, the central tendency (the median speed) for each of the tasks ranged from about 30 (letter sound knowledge) to about 80 (passage reading) items per minute. It shows that the students had more fluency with reading connected words (passage reading) than with phonics.

The variation (range of scores) for each of the tasks varied from about 130 (letter sound knowledge) to about 180 (passage reading). It shows that the scores were more spread out when reading connected words than providing the sound of letters distributed in a random order.

Note also that the medians and the ranges increased from grade 3 to grade 5 for all fluency tasks. Many students are becoming more fluent readers at grade 5, but there are also those students who are either non-readers or very low readers. These children lack of knowledge of letter names, sight words, connected text, and (especially) phonics.

FIGURE A3: PHONICS AND READING-RATE FLUENCY BOX PLOTS FOR GRADE 5



Annex 3: Examples of Fluency Score Threshold Calculations

There are different ways of interpreting test scores. Three of the main ways are 1) raw scores (e.g., number correct), 2) scale scores (e.g., percent correct), and 3) percentile scores (e.g., rank in relation to other students). In the report, we presented scores in terms of number correct (for the fluency tasks) and percent correct (for all tasks). We could also calculate the percentile scores for each student, though this is not normally done with EGRA. Note that these kinds of calculations do not change or affect the actual results, but they do involve issues of interpretability.

A fourth main way of interpreting scores is through performance categories, e.g., low, middle, and high. This requires setting cut-scores, or thresholds, to separate the student scores into categories, e.g., two cut-scores lead to three performance categories. The following analysis shows two examples of calculating thresholds for passage reading scores (CWPM), which allows us to place the student scores into different performance categories. Note that performance categories are often accompanied by performance level descriptors (PLDs), which give a text-based explanation of the meaning of the scores in each category. We have not developed PLDs for these examples since 1) the threshold setting is at a preliminary stage and 2) reading specialists with knowledge of local curricula and context generally develop the PLDs.

Fluency using an 80 percent comprehension threshold

In the first example, we used a method that has been suggested by some EGRA specialists. It involves calculating the mean reading speed associated with 80 percent comprehension for those that can read at least one word correctly and then applying it as a fluent cut-score. In other words, the mean reading speed for these students signifies whether the students are fluent readers through using both passage reading speed *and* comprehension in the calculation; the fluent cut-score separates the fluent readers from the non-fluent readers. To establish a second threshold, we again followed the suggested method and used the lowest level of reading (1 CWPM) as the non-fluent cut-score. The two cut-scores resulted in three performance levels: non-readers (low), non-fluent readers (middle), and fluent readers (high).

At grade 3, the mean reading speed on the passage reading task (Task 7a) for students who scored 80 percent on the passage comprehension task (Task 7b) was 89.1 (rounded to 89). With this method, 89 CWPM becomes a threshold for grade 3 students who are proficient at passage reading *and* comprehension. At grade 5, the mean speed on the passage reading task (Task 7a) for students who scored 80 percent on the passage comprehension task (Task 7b) was 98.2 (rounded to 93). Then 93 CWPM becomes a threshold for grade 5 students who are proficient at passage reading *and* comprehension. CWPM

The definitions of the three categories in terms of CWPM and the percentages of grades 3 and 5 students in the categories are shown in Table A2 below.

TABLE A2: THRESHOLDS FOR CWPM WITH 80 PERCENT COMPREHENSION

Category (Performance Level)	Grade 3		Grade 5	
	CWPM	% of Students	CWPM	% of Students
Non-Reader	0	22.7%	0	2.9%
Non-Fluent Reader	1 to 88	73.1%	1 to 97	65.7%
Fluent Reader	89 and above	4.2%	98 and above	31.4%
Total	--	100.0%	--	100.0%

Note that the majority of the students are in the middle category at each grade level. This is due the large range of scores for this category, i.e., from the students who score just above non-readers to those who score just below fluent readers are in the non-fluent reader (middle) category.

Fluency using fixed interval thresholds

In the second example, we used fixed intervals of CWPM for the performance levels. This reduced the problem of having a large range of students in the middle category by creating early reader and intermediate reader categories. It also follows common practice when setting performance categories of having between three and five levels for student scores. We used an interval of 40 CWPM to produce five performance levels, along with a category for the non-readers. The five levels were: non-readers (0 CWPM); early readers (1-40 CWPM); intermediate readers (41-80 CWPM); fluent readers (81-120 CWPM); and advanced readers (121 and above CWPM).

TABLE A3: THRESHOLDS FOR CWPM WITH FIXED INTERVALS

Category (Performance Level)	CWPM	% of Students	
		Grade 3	Grade 5
Non-Reader	0	22.7%	2.9%
Early Reader	1 to 40	45.1%	12.6%
Intermediate Reader	41 to 80	26.3%	37.4%
Fluent Reader	81 to 120	5.1%	34.5%
Advanced Reader	121 and above	0.8%	12.7%
Total	--	100.0%	100.0%

At both grades 3 and 5, the fixed interval method allowed for more distribution of the scores across the categories. We can also see a shift in percentages of students in each category from grade 3 to grade 5; the performance categories allow for a score interpretation showing that students are improving across the grade levels, with more scores in the lower categories at grade 3 and more scores in the higher categories at grade 5.

Remarks

While it is possible to use such percentages to set cut-scores for interpretation purposes at the baseline, midline and endline, this analysis should be taken as preliminary. For instance, more well-known and accepted method of setting thresholds – which is commonly called "standard setting" by measurement specialists – involve holding a workshop with local reading experts to set the cut-scores according to the experts' conceptions of what students should know and be able to do in order to be classified into a performance category. There are several well-known methods, e.g., Angoff and Bookmark, which have been judged as valid and reliable for this purpose.[4] Further discussions on setting thresholds involving local reading experts are recommended.

[4] References include: Zieky, M. & Perie, M. (2006). *A primer on setting cut-scores on tests of educational achievement*. Princeton, New Jersey: Educational Testing Service; Cizek, G. (1996). *Standard-setting guidelines*. Educational Measurement: Issues and Practices, Spring 1996, p. 13-21; Cizek, G., Bunch, M., & Koons, H. (2004). *Setting performance standards: Contemporary methods*. Educational Measurement: Issues and Practices, Winter 2004.

Annex 4: Distribution of Reading Fluency and Comprehension Scores using Fixed Intervals

In this last annex, we provide more information on the relationship between reading fluency (speed) and comprehension using information from the fixed interval method. While the data show a positive relationship between speed and comprehension, there are sizeable numbers of "fluent" readers with little comprehension. Our conclusion is that setting a cut-score using a less than reliable indicator, such as the mean speed of students with 80 percent comprehension (i.e., using *both* speed and comprehension), can be problematic. The result is categorizing some students as fluent readers who in fact, according to the definition, are not, i.e., they have high reading speed but low comprehension. It may be better to set thresholds based solely on a single indicator – reading speed – rather than mixing it with comprehension.

The figures and tables below (Tables A4-A5 and Figures A4-A5) expand on the data in Table A3. They show the results for reading fluency (in terms of speed) by comprehension level for grades 3 and 5. We used the categories based on intervals of 40 CWPM, along with a category for the CWPM non-readers (0 CWPM). Comprehension levels were calculated in terms of percent correct scores (e.g., 20 percent is the same as correctly answering one question out of five total questions). For instance, at grade 3, 100 percent of the non-readers have 0 percent comprehension and 14 percent of the advanced readers have 80 percent comprehension.

TABLE A4: GRADE 3 READING FLUENCY AND COMPREHENSION

Category	CWPM	% of Students by Comprehension Level						
(Performance Level)		0%	20%	40%	60%	80%	100%	Total
Non-Reader	0	100%	0%	0%	0%	0%	0%	100%
Early Reader	1 to 40	85%	12%	2%	1%	0%	0%	100%
Intermediate Reader	41 to 80	43%	26%	15%	15%	1%	0%	100%
Fluent Reader	81 to 120	14%	11%	20%	25%	25%	5%	100%
Advanced Reader	121 and above	43%	0%	14%	29%	14%	0%	100%

FIGURE A4: GRADE 3 READING FLUENCY AND COMPREHENSION

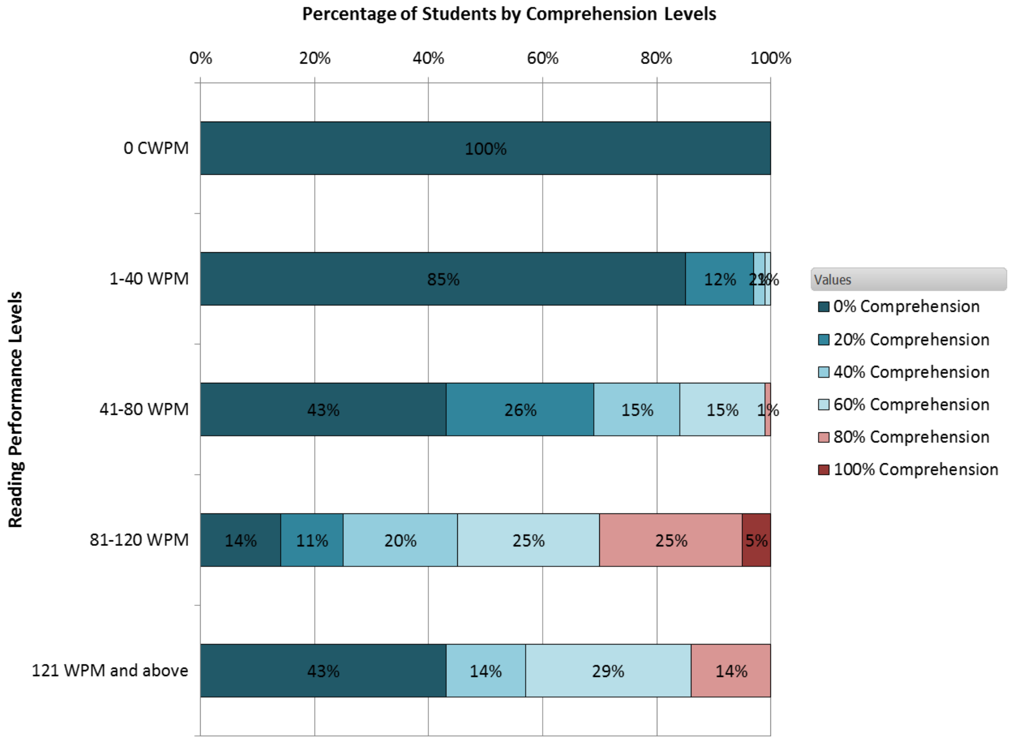
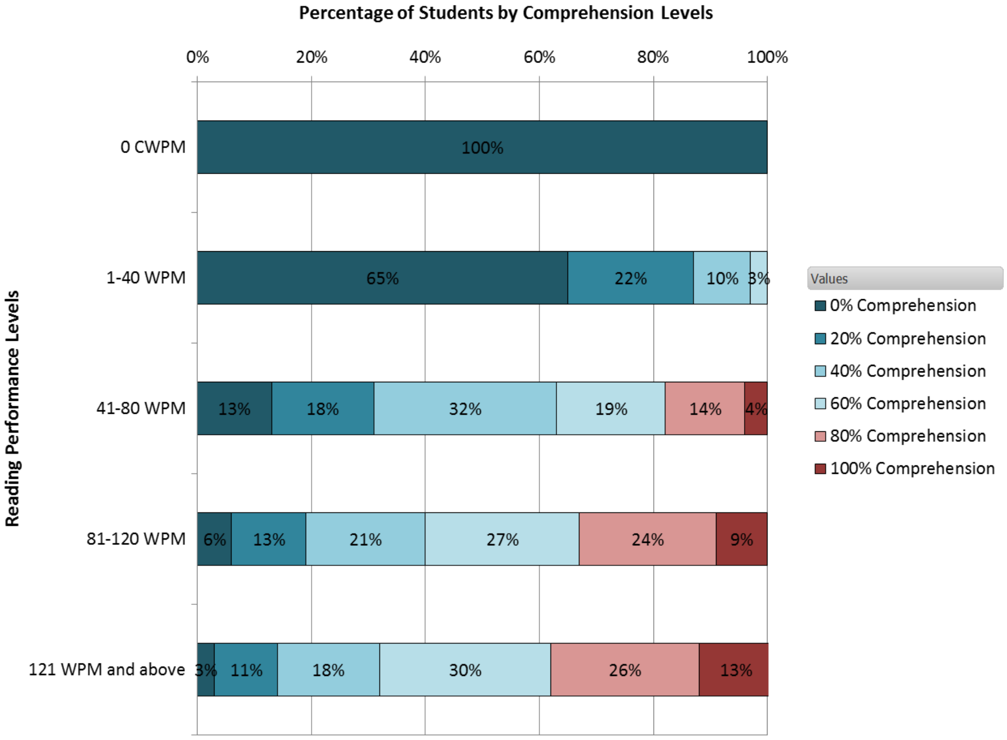


TABLE A5: GRADE 5 READING FLUENCY AND COMPREHENSION

Category (Performance Level)	CWPM	% of Students by Comprehension Level						
		0%	20%	40%	60%	80%	100%	Total
Non-Reader	0	100%	0%	0%	0%	0%	0%	100%
Early Reader	1 to 40	65%	22%	10%	3%	0%	0%	100%
Intermediate Reader	41 to 80	13%	18%	32%	19%	14%	4%	100%
Fluent Reader	81 to 120	6%	13%	21%	27%	24%	9%	100%
Advanced Reader	121 and above	3%	11%	18%	30%	26%	13%	100%

FIGURE A5: GRADE 5 READING FLUENCY AND COMPREHENSION



The main results for the categories of reading speed (from non-readers to advanced readers) in relation to comprehension levels (from 0 percent to 100 percent) for grades 3 and 5 are summarized as follows:

- Non-Readers (0 CWPM) – All of the non-readers had 0 percent comprehension.
- Early Readers (1-40 CWPM) – Most of the early readers (85 percent at grade 3 and 65 percent at grade 5) had 0 percent comprehension. None of them achieved 80 percent comprehension.
- Intermediate Readers (41-80 CWPM) –43 percent at grade 3 and 13 percent at grade 5 had 0 percent comprehension. A small minority of them (1 percent at grade 3 and 18 percent at grade 5) achieved at least 80 percent comprehension.
- Fluent Readers (81-120 CWPM) – About 14 percent of the fluent readers at grade 3 and 6 percent at grade 5 had 0 percent comprehension. Around one-third of them (30 percent at grade 3 and 33 percent at grade 5) achieved at least 80 percent comprehension.
- Advanced Readers (121 CWPM and above) – Only a minority of the advance readers (14 percent at grade 3 and 39 percent at grade 5) read with at least 80 comprehension.

The key point from the data is that most of the fluent and advanced readers – at both grade levels – did not reach 80 percent comprehension. Setting a threshold under the assumption that fluent readers (in terms of speed) have a high level of comprehension can be misleading. Conversely, using a single indicator, i.e., reading speed, to set thresholds can be a more reliable way of interpreting the results.